On the role of weights rounding in applications of resampling based on pseudo-populations

F. Andreis P.L. Conti F. Mecatti

Abstract

Resampling methods are widely studied and increasingly employed in applied research and practice. When dealing with complex sampling designs, common resampling techniques require to adjust non-integer sampling weights in order to construct the so called "pseudo-population" where to perform the actual resampling. The practice of rounding, however, has been empirically shown to be harmful under general designs. In this paper we present asymptotic results concerning, in particular, the practice of rounding resampling weights to the nearest integer, an approach that is commonly adopted by virtue of its reduced computational burden, as opposed to randomization-based alternatives. We prove that such approach leads to non-consistent estimation of the distribution function of the survey variable; we provide empirical evidence of the practical consequences of the non-consistency when point estimation of the variance of complex estimators is of interest.
KEYWORDS: π-ps complex designs - bootstrap - probability proportional to size - finite populations - variance estimation

1 Introduction

Resampling is a popular computer intensive tool for assessing estimators accuracy, constructing confidence intervals and computing p-values. For general sampling designs, such as probability proportional to size (π-ps) designs, the need arises to produce adaptations of the classic Bootstrap approach to account for the non-*iid* nature of the sample data. Many proposals have appeared in the literature, based on using weighting systems in the resampling and/or in the estimation procedure (see, e.g., [Antál and Tillé (2011)], [Beaumont and Patak (2012)], [Ranalli and Mecatti (2012)]). The use of integer weights would guarantee desirable analytical properties of both the resampling procedure and the final Bootstrap estimates, but this does not usually occur in real applications. The main suggestions to bypass the non-integer weights issue include i) randomization and ii) systematical rounding. Although the general opinion in the literature is that these approaches have little effect on the resampling procedures and on the quality of the estimates, it has been noted that both solutions affect them to a non-negligible extent. This has been empirically investigated in [Andreis and Mecatti (2015)], where the authors conclude, on the basis of an extended simulation study, that even in the simple case of estimating the variance of the Horvitz-Thompson estimator for the mean, rounding can detrimentally affect the final Bootstrap estimates properties. This

paper aims at providing a theoretical support to empirical evidence, as well as to extend the investigation of the rounding effect to more complex semi- or non-linear estimators.

Section 2 discusses the main assumptions and states some preliminary results. Section 3 outlines a general Bootstrap algorithm applying to non-iid sample data where randomization is employed, and provides the first asymptotic results, that will then be compared in Section 4 to those concerning the systematical rounding approach. Section 5 provides empirical evidence concerning the effect of rounding on basic Bootstrap mimicking principles and on the estimation of relevant finite population quantities. Section 6 contains the final remarks for this paper.

2 Basic assumptions and preliminary results

Consider a finite population $\mathscr{U}_N$ of size N; a sample $\mathbf{s}$ of size $n_\mathbf{s}$ is a subset $\mathbf{s}$ of $n_\mathbf{s}$ units of $\mathscr{U}_N$. For each unit $i \in \mathscr{U}_N$, define a Bernoulli random variable (r.v.) D_i, such that $D_i = 1$ if i is included in $\mathbf{s}$, 0 otherwise and let $\mathbf{D}_N = (D_1, \ldots, D_N)$. Of course, $n_\mathbf{s} = D_1 + \cdots + D_N$.

A sampling design P is the probability distribution of $\mathbf{D}_N$. The expectations w.r.t. the sampling design P $\pi_i = E_P[D_i]$ and $\pi_{ij} = E_P[D_i D_j]$ are the first and second order inclusion probabilities, respectively. In view of their importance, we will confine ourselves to fixed size sampling designs: $n_s \equiv n$.

In π-ps sampling designs, the first order inclusion probabilities are chosen to be proportional to an auxiliary variable $\mathscr{X}$, traditionally a measure of size known for each population unit, *i.e.* $\pi_i \propto x_i$, $i = 1, \ldots, N$; cfr., e.g., [Hájek (1981)], [Tillé (2006)].

If the r.v.s D_is are independent with expectations $\pi_i \in (0,1)$, then the corresponding design is the *Poisson sampling design*. The *rejective sampling* (or *normalized conditional Poisson sampling*, cfr. [Hájek (1964)], [Tillé (2006)]) is obtained from the Poisson sampling by conditioning w.r.t. $n_s = n$. In the sequel, the latter will be denoted by the suffix R. The *Hellinger distance* between a sampling design P and the rejective design is defined as

$$d_H(P, P_R) = \sum_{i=1}^{N} \left(\sqrt{P_P(D_i)} - \sqrt{P_R(D_i)} \right)^2 .$$

Let y_i (x_i) be the value of the study variable $\mathscr{Y}$ (auxiliary variable $\mathscr{X}$) for unit i. The assumptions on which the present paper rests are similar to those in [Conti et al. (2015)]. They are listed below.

A1. $(\mathscr{U}_N;\, N \geq 1)$ is a sequence of finite populations of increasing size N.

A2. For each N, (y_i, x_i), $i = 1, \ldots, N$, are ~~are~~ realizations of a superpopulation $\{(Y_i, X_i),\ i = 1, \ldots, N\}$ of *i.i.d.* two-dimensional r.v.s, with probability distribution $\mathbb{P}$. The (superpopulation) distribution function (df) of (Y_i, X_i) is denoted by

$$H(y, x) = \mathbb{P}(Y_i \leq y, X_i \leq x)$$

and the corresponding marginal dfs of Y_i and X_i by

$$F(y) = \mathbb{P}(Y_i \leq y), \;\; G(x) = \mathbb{P}(X_i \leq x),$$

respectively.

A3. For each population $\mathscr{U}_N$, sample units are selected according to a fixed size sample design P with first order inclusion probabilities π_is proportional to x_is, and sample size n. Furthermore,

$$d = \mathbb{E}[\pi_i(1-\pi_i)] \tag{1}$$

is assumed to be positive.

A4. The sample size n increases as the population size N does, with

$$\lim_{N\to\infty} \frac{n}{N} = f, \;\; 0 < f < 1.$$

A5. For each population $(\mathscr{U}_N;\, N \geq 1)$, let P_R be the rejective sampling design with inclusion probabilities $\pi_1, \ldots, \pi_N$, and let P be the actual sampling design (with the same inclusion probabilities). Then

$$d_H(P, P_R) \to 0 \;\text{ as } N \to \infty.$$

The possible dependence between the study variable and the auxiliary variable is addressed by assumptions A2, A3. Notice that a positive correlation (approximate proportionality) is the basic motivation for choosing a π-ps sampling design; however, no assumption is made on such a dependence, apart from its possible existence. The sampling designs satisfying assumption A5 are essentially designs with asymptotically maximal entropy such as, for example, the Rao-Sampford, Chao, and successive sampling designs (cfr.[Berger (1998)], [Berger (2011)], [Conti (2014)]); an example of design for which A5 does not hold is the non-randomized systematic sampling.

2.1 Population parameters and statistical functionals

Let $F_N(y) = N^{-1}\sum_{i=1}^N I_{(y_i \leq y)}$ be the population distribution function (p.d.f.), where $I_{(y_i \leq y)}$ is equal to 1 iff $y_i \leq y$, and 0 otherwise. A *finite population parameter* is a functional $\theta_N = \theta(F_N)$ of the df. In order to estimate a parameter of this form, a natural approach consists in estimating first F_N by an appropriate estimator $\widehat{F}$, and then computing $\widehat{\theta} = \theta(\widehat{F})$. This is the classical approach of *statistical functionals* (cfr., for instance, [Serfling (1980)]). In case of *i.i.d.* observations, the commonly employed estimator $\widehat{F}$ is the empirical distribution function. In the present case, where a general (possibly complex) sampling design is used, as a natural estimator of F_N we consider the *Hájek estimator* $\widehat{F}_H$, i.e. $\widehat{F}_H(y) = \sum_{i\in\mathbf{s}} \pi_i^{-1} I_{(y_i \leq y)} / \sum_{i\in\mathbf{s}} \pi_i^{-1}$. On the basis of the approach outlined above, the finite population parameter θ_N is then estimated by $\widehat{\theta}_H = \theta\left(\widehat{F}_H\right)$.

2.2 Preliminary results

The asymptotic properties of the estimators $\widehat{F}_H(y)$ and $\widehat{\theta}_H$, as both the sample size and the population size increase, are studied in [Conti et al. (2015)]. As far as the functional $\theta(F_N)$ is concerned, the key assumption is its *Hadamard-differentiability* (cfr. [van der Vaart (1998)]); we hereafter denote by $\theta'_F(\cdot)$ the *Hadamard derivative* of θ at F. Commonly encountered examples of functionals that fall in this class include the moments and the quantiles; moreover, linear combinations and ratios thereof are also Hadamard-differentiable. The main asymptotic results in [Conti et al. (2015)] can be summarized by the following proposition.

Proposition 1
Consider the stochastic process $W_N^H(\cdot) = (W_n^H(y);\ y \in \mathbb{R})$, where

$$W_N^H(y) = \sqrt{n}(\widehat{F}_H(y) - F_N(y));\ y \in \mathbb{R}, \tag{2}$$

and suppose that assumptions A1-A5 are fulfilled with F continuous, and that $\theta(\cdot)$ is (continuously) Hadamard-differentiable at F tangentially to the set of continuous functions on $[a, b] \subseteq \mathbb{R}$, with Hadamard derivative $\theta'_F(\cdot)$. The following two statements hold.

(*i*) With $\mathbb{P}$-probability 1, conditionally on $\mathbf{y}_N$, $\mathbf{x}_N$ the sequence $W_N^H(\cdot),\ N \geq 1$, converges weakly, in $D[-\infty, +\infty]$ equipped with the Skorokhod topology, to a Gaussian process $W^H(\cdot) = (W^H(y);\ y \in \mathbb{R})$ with zero mean function, and covariance kernel

$$\begin{aligned} C^H(y,t) &= f\left\{\frac{\mathbb{E}[X_1]}{f}K_{-1}(y\wedge t) - 1\right\}F(y\wedge t) \\ &\quad -\frac{f^3}{d}\left(1-\frac{K_1(y)}{\mathbb{E}[X_1]}\right)\left(1-\frac{K_1(t)}{\mathbb{E}[X_1]}\right)F(y)F(t) \\ &\quad -f\left\{\frac{\mathbb{E}[X_1]}{f}\left(K_{-1}(y)+K_{-1}(t)-\mathbb{E}\left[X_1^{-1}\right]-1\right)\right\} \\ &\times\ F(y)F(t), \end{aligned} \tag{3}$$

with d given by (1), and $K_\alpha(y) = \mathbb{E}\left[X_1^\alpha \,\middle|\, Y_1 \leq y\right], y \in \mathbb{R}, \alpha = \pm 1$.

(*ii*) With $\mathbb{P}$-probability 1, and conditionally on $\mathbf{y}_N$, $\mathbf{x}_N$, the sequence of functionals $(\sqrt{n}(\theta(\widehat{F}_H(y)) - \theta(F_N));\ N \geq 1)$ converges weakly to $\theta'_F(W^H)$, as N increases.

In Proposition 1 the actual population y_is and x_is values are considered as *fixed*. The only source of variability is the sampling design. If we let the population size N go to infinity, we must also consider corresponding sequences $\mathbf{y}_\infty = (y_1, y_2, \dots)$, $\mathbf{x}_\infty = (x_1, x_2, \dots)$ of y_is and x_is values. The actual $\mathbf{y}_N = (y_1, \dots, y_N)$, $\mathbf{x}_N = (x_1, \dots, x_N)$ are the segments of the first N y_is, x_is in the sequences $\mathbf{y}_\infty$, $\mathbf{x}_\infty$, respectively. As N increases, $\mathbf{y}_N$ tends to $\mathbf{y}_\infty$ and $\mathbf{x}_N$ tends to $\mathbf{x}_\infty$.

3 Resampling based on pseudo-population: theoretical properties

3.1 The resampling algorithm

In this section, a short description of the resampling procedure for π-ps designs, as proposed in [Holmberg (1998)] and based on the Bootstrap population approach introduced in [Gross (1980)], [Chao and Lo (1985)], is provided. Let $\mathbf{s}$ be the sample selected from $\mathscr{U}_N$, and, for each unit i in $\mathbf{s}$, let $\lfloor \pi_i^{-1} \rfloor$ be the largest integer smaller than π_i^{-1}, and $r_i = \pi_i^{-1} - \lfloor \pi_i^{-1} \rfloor$. Define further, for each $i \in \mathbf{s}$, independent Bernoulli r.v.s ε_i such that $Pr(\varepsilon_i = 1|\mathbf{s}) = r_i$, $Pr(\varepsilon_i = 0|\mathbf{s}) = 1 - r_i$.

1. *For every $i \in \mathbf{s}$, generate n independent Bernoulli r.v.s ε_i defined as below, and set $T_i = \lfloor \pi_i^{-1} \rfloor + \varepsilon_i$*

2. *Define $N^* = \sum_{i \in \mathbf{s}} T_i$, and construct a "pseudo-population" $\mathscr{U}_{N^*}^*$ of size N^*, where $i \in \mathbf{s}$ is copied T_i times. The T_i units $k \in \mathscr{U}_{N^*}^*$ that are copies of $i \in \mathbf{s}$ are assigned values $(x_k^*, y_k^*) = (x_i, y_i)$. Let*

$$F_{N^*}(y) = \frac{1}{N^*} \sum_{k=1}^{N^*} I_{(y_k^* \leq y)}, \;\; y \in \mathbb{R} \tag{4}$$

   be the df of the pseudo-population.

3. *Select a sample $\mathbf{s}^*$ of n units from $\mathscr{U}_{N^*}^*$, by applying the same sampling scheme used to draw $\mathbf{s}$ from $\mathscr{U}_N$. This implies than the inclusion probabilities π_k^*, $k \in \mathscr{U}_{N^*}^*$, are equal to*

$$\pi_k^* = n \frac{x_k^*}{\sum_{k=1}^{N^*} x_k^*} = n \frac{x_k^*}{\sum_{i \in \mathbf{s}} T_i x_i} \tag{5}$$

4. *Consider the Hájek estimator $F_H^*(y)$ of $F_{N^*}(y)$ constructed on the basis of the sample $\mathbf{s}^*$ of the pseudo-population, and let $\widehat{\theta}_H^* = \theta(F_H^*)$ be the corresponding estimator of $\theta(F_{N^*})$. The resampling principle is based on a simple idea:* the (design-based) probability law of $\widehat{\theta}_H = \theta(\widehat{F}_H)$, given $\mathbf{x}_N$, $\mathbf{y}_N$, is approximated by the (resampling-based) probability law of $\widehat{\theta}_H^* = \theta(F_H^*)$, given $\mathbf{x}_N$, $\mathbf{y}_N$, $\mathbf{s}$.

3.2 Properties of the resampling procedure based on pseudo-population

The main result of the present section is as follows: as the sample and the population size become large, the pseudo-population $\mathscr{U}_{N^*}^*$ becomes "similar" to the actual population $\mathscr{U}_N$. As a consequence, the (resampling) probability law of $\hat{\theta}_H^*$ becomes similar to the (sampling) probability law of $\hat{\theta}_H$. More formally, as shown in [Conti et al. (2015)], the pseudo-population size is asymptotically equivalent to the population size:

$$\frac{N^*}{N} \to 1 \tag{6}$$

in probability as $N \to \infty$.

In the second place, the resampling procedure at steps 1-4 essentially produces a "resampled version" of the process (2), *i.e.*

$$W_N^{H*}(y) = \sqrt{n}(\widehat{F}_H^*(y) - F_{N^*}(y)),\ y \in \mathbb{R};\ N \geq 1. \tag{7}$$

In Proposition 2 (cfr. [Conti et al. (2015)]), it is established that the (sampling design based) asymptotic probability law of W_N^H coincides with the (resampling based) probability law of W_N^{H*}.

Proposition 2
Under the assumptions of Proposition **1**, and if the resampling design satisfies assumptions A1-A5, conditionally on $\mathbf{y}_N$, $\mathbf{x}_N$, $\mathbf{D}_N$, ε_is, as N goes to infinity the following two statements hold true.

- (i) The sequence of stochastic processes $(W_N^{H*}(\cdot);\ N \geq 1)$, converges weakly, in $D[-\infty, +\infty]$ equipped with the Skorokhod distance, to a Gaussian process $W^H(\cdot)$ with zero mean function and covariance kernel (3).
- (ii) $(\sqrt{n}(\theta(\widehat{F}_H^*(y)) - \theta(F_N^*));\ N \geq 1)$ converges weakly to $\theta_F'(W^H)$.

Note that the assumption of identical sampling and resampling design, as requested in [Holmberg (1998)], is here relaxed: the results in Proposition 2 remain valid also if the sampling design and the resampling design differ. The key points are actually three. First of all, the size N^* of the pseudo-population asymptotically behaves (in probability) as the size N of the "actual" population, as in $N^*/N \xrightarrow{p} 1$ as $N \to \infty$. In the second place, the pseudo-population itself, as N increases, becomes increasingly "similar" to the original one. Finally, the first order inclusion probabilities in the resampling design are chosen similarly to those of the sampling design acting on the "original" population.
For computational reasons the resampling probability distribution of $\widehat{F}_H^*(y)$ is usually approximated via Monte Carlo simulation. Steps 3 and 4 are repeated M times, so that M independent replicates $\widehat{\theta}_{H,m}^*$, $m = 1, \ldots, M$ are generated, with M large enough. Originally, Holmberg's proposal (as well as almost all papers devoted to resampling from finite populations) is essentially devoted to estimate the variance of $\widehat{\theta}_H$, and not the whole distribution. Here we mainly focus on the approximation of the distribution of the estimator $\widehat{\theta}_H$, according to a basic Bootstrap principle, cfr. for instance [Hall (1992)].

4 Simplified resampling based on rounding weights

4.1 The simplified resampling algorithm

Since the weights π_i^{-1} are hardly ever integer, the unconditional resampling scheme of Sub-section 3.1 is usually cumbersome from a computational point of view. For this reason, a "simplified version" of the resampling scheme of Sub-section 3.2, based on rounding the weights π_i^{-1} has been proposed in the literature (see, for instance,

[Chauvet (2007)], [Andreis and Mecatti (2015)] and references therein). With the same notation as in Section 3.1, define

$$\widetilde{T}_i = \lfloor \pi_i^{-1} \rfloor + I_{(r \geq 1/2)} \tag{8}$$

where $I_{(r \geq 1/2)}$ is 1 iff $r \geq 1/2$, and 0 otherwise. The simplified resampling algorithm based on systematically rounded weights is similar to the resampling algorithm in Subsection 3.1, with T_i replaced by $\widetilde{T}_i, i = 1, ..., N$. All relevant quantities are now defined in terms of the new weights and are denoted by $\widetilde{N}^*$, $F_{\widetilde{N}^*}(y)$, $\widetilde{\pi}_k^*$, $\widetilde{F}_H^*(y)$ and $\widetilde{\theta}_H^*$, respectively.

4.2 What is the effect of rounding weights? Theoretical results

The intuition behind the simplified version of the resampling scheme is that the rounding errors arising in replacing π_i^{-1} with $\lfloor \pi_i^{-1} \rfloor + I_{(r \geq 1/2)}$ should compensate, so that, when both the sample size and the population size are large enough, they should have a negligible effect. In fact, this is often claimed when bootstrapping non-iid sample (see for instance [Beaumont and Patak (2012)]). However, this intuition is wrong, because, even under favourable conditions, the effect of rounding errors does not generally disappear in asymptotic analysis. The effect of rounding is evaluated in the sequel.

From the strong law of large numbers, and by assumption $A4$, as N gets large,

$$f_N = \frac{n}{N} \approx f, \;\; \frac{1}{N} \sum_{i=1}^{N} x_i \approx \mathbb{E}[X_1]$$

with $\mathbb{P}$-probability 1. Hence, to simplify the reasoning, we will take

$$U_i \approx \frac{\mathbb{E}[X_1]}{f} X_i^{-1} = \pi_i^{-1}, \;\; V_i = \lfloor U_i \rfloor, \;\; R_i = U_i - V_i. \tag{9}$$

The r.v.s V_is are *i.i.d.*, and take positive integer values. In the sequel, we will use the notation:

$$p_k = \mathbb{P}(V_i = k), \;\; k \geq 1. \tag{10}$$

The r.v.s R_is are *i.i.d.*. We will further assume that they possess a density function, and we will denote by $g_k(r)$ the density of R_i conditionally on $V_i = k$, $0 < r < 1$. As already said, the idea behind rounding is that the rounding errors compensate to some extent. The simplest (and most favourable, in many respects) way to formalize this intuition consists in assuming that $g_k(r)$ is symmetric around $1/2$:

$$g_k(r) = g_k(1-r), \;\; 0 < r < 1/2. \tag{11}$$

Proposition 3
Assume that assumptions A1-A5 hold, and let $\widetilde{N}^* = \sum \widetilde{T}_i D_i$ be the pseudo-population size under the rounding scheme described so far, Then:

$$\frac{\widetilde{N}^*}{N} \xrightarrow{p} 1 + c \neq 1, \;\; \text{as } N, n \to \infty \tag{12}$$

Furthermore, if (11) holds, then $c < 0$.
Proof - From the law of large numbers, $\widetilde{N}/N$ tends in probability to

$$\begin{aligned}
& \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\widetilde{T}_i E\left[D_i \,|\, \mathbf{x}_N, \mathbf{y}_N\right]\right] \\
= \;& \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\pi_i \widetilde{T}_i\right] \\
= \;& \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\pi_i(\pi_i^{-1} - R_i I_{(R_i<1/2)} + (1-R_i)I_{(R_i\geq 1/2)})\right] \\
= \;& 1 + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\pi_i(I_{(R_i\geq 1/2)} - R_i)\right] \\
= \;& 1 + \mathbb{E}\left[U_1^{-1}(I_{(R_1\geq 1/2)} - R_1)\right]. \qquad (13)
\end{aligned}$$

The expectation at the right hand side of (13) is equal to

$$\begin{aligned}
& \sum_{k=1}^{\infty} p_k \left\{ \int_0^1 \frac{1}{k+r}(I_{(r\geq 1/2)} - r) g_k(r) dr \right\} \\
= \;& \sum_{k=1}^{\infty} p_k \left\{ -\int_0^{1/2} \frac{r}{k+r} g_k(r) dr + \int_{1/2}^1 \frac{1-r}{k+r} g_k(r) dr \right\} \\
= \;& \sum_{k=1}^{\infty} p_k \left\{ -\int_0^{1/2} \frac{r}{k+r} g_k(r) dr + \int_0^{1/2} \frac{t}{k+1-t} g_k(1-t) dt \right\} \\
= \;& c
\end{aligned}$$

and this quantity is not generally equal to zero. Hence, we have:

$$\frac{\widetilde{N}^*}{N} \xrightarrow{p} 1 + c \;\text{ as } N \to \infty$$

with $c \neq 0$. In particular, in the best scenario (11), we have further

$$\begin{aligned}
& \mathbb{E}\left[U_1^{-1}(I_{(R_1\geq 1/2)} - R_1)\right] = \qquad (14) \\
& \sum_{k=1}^{\infty} p_k \left\{ \int_0^{1/2} r \left(\frac{1}{k+1-r} - \frac{1}{k+r} \right) dr \right\} < 0
\end{aligned}$$

because $(k+1-r)^{-1} < (k+r)^{-1}$ for every $0 < r < 1/2$. Hence, $\widetilde{N}^*/N$ tends in probability to a constant which is *strictly smaller than* 1.

To better understand the effect of rounding, it is worth studying the different behaviour of $F_{N^*}(y)$ and $F_{\widetilde{N}^*}(y)$, as N increases. As an easy consequence of the law of large numbers (or, as an alternative, of Propositions **1** and **2**), $F_N^*(y)$ tends in probability to $F(y)$ as N increases; in symbols:

$$F_{N^*}(y) \xrightarrow{p} F(y) \;\text{ as } N \to \infty. \qquad (15)$$

The same *does not hold* under the rounding scheme.

Proposition 4
Under the assumptions of Proposition **3**, $F_{\widetilde{N}^*}(y)$ tends in probability to

$$\frac{1+c(y)}{1+c}F(y) \neq F(y) \tag{16}$$

as N increases, with $[1+c(y)]/(1+c) \neq 1$.
Proof - Proposition **3** and the law of large numbers imply that $F_{\widetilde{N}^*}(y)$ tends in probability to

$$(1+c)^{-1}\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N} E\left[D_i\widetilde{T}_i \,\middle|\, \mathbf{x}_N, \mathbf{y}_N\right] I_{(y_i \leq y)}\right]. \tag{17}$$

Taking into account that

$$E\left[D_i\widetilde{T}_i \,\middle|\, \mathbf{x}_N, \mathbf{y}_N\right] = \pi_i\left(\pi_i^{-1} + I_{(R_I \geq 1/2)} - R_i\right)$$

the expectation in Equation (17) is equal to

$$(1+c)^{-1}\left(F(y) + \mathbb{E}\left[\pi_i(I_{(R_I \geq 1/2)} - R_i)I_{(Y_i \leq y)}\right]\right). \tag{18}$$

Next, if $g_k(r|Y=y)$ ($g_k(r|Y \leq y)$) denotes the density function of R_i given $V_i = k$ and $Y_i = y$ ($V_i = k$ and $Y_i \leq y$), and if $p_k^t = \mathbb{P}(V_i = k|Y_i = t)$ ($p_k^{Y \leq y} = \mathbb{P}(V_i = k|Y_i \leq y)$), the expectation at the right hand side of Equation (18) is equal to

$$\begin{aligned}
&\mathbb{E}\left[I_{(Y_i \leq y)}\mathbb{E}\left[U_i^{-1}(I_{(R_i \geq 1/2)} - R_i)\,\middle|\, Y_i = t\right]\right] \qquad (19)\\
&= \int_{-\infty}^{y}\left\{\sum_{k=1}^{\infty} p_k^t \int_0^1 \frac{1}{k+r}\left(I_{(r \geq 1/2)} - r\right) g_k(r|t)dr\right\} dF(t)\\
&= \left\{\sum_{k=1}^{\infty} p_k^{Y \leq y} \int_0^1 \frac{1}{k+r}\left(I_{(r \geq 1/2)} - r\right) g_k(r|Y \leq y)dr\right\} F(y).
\end{aligned}$$

In general, the quantity

$$c(y) = \sum_{k=1}^{\infty} p_k^{Y \leq y} \int_0^1 \frac{1}{k+r}\left(I_{(r \geq 1/2)} - r\right) g_k(r|Y \leq y)dr \tag{20}$$

depends on y, and tends to c as y goes to ∞. At any rate, from (18) we get (16).

Since Proposition **2** heavily depends on the validity of Equation (6), it is apparent that it does not hold anymore. From Equation (16) it can be argued that the the sequence of stochastic processes

$$\widetilde{W}_N^{H*}(y) = \sqrt{n}(\widetilde{F}_H^*(y) - F_{\widetilde{N}^*}(y)),\ y \in \mathbb{R};\ N \geq 1$$

converges weakly to a Gaussian process with zero mean function and covariance kernel of the form (3), but with $F(\cdot)$ replaced by $[1+c(y)]/(1+c)F(\cdot)$. In other terms, the simplified resampling does not asymptotically reproduce the (design based) probability distribution of $\widehat{F}_H(y)$. The same conclusion, of course, holds for general functionals of the form $\theta(\widetilde{F}_H^*)$.

5 Some empirical evidence

In this Section, we present simulation results aimed at investigating two aspects of the problem at hand: i) how well do the pseudo-populations based on the two rounding approaches manage to reproduce the original population, and ii) what is the effect of the choice of rounding method on a typical application of Bootstrap, specifically the estimation of the variance of a complex functional of the survey variable. All computations and plots have been produced within the R environment ([R Core Team (2015)]).

5.1 Mimicking of the original population

Following the fundamental Bootstrap principle of mimicking, the pseudo-populations constructed using the methods described in paragraphs 3.1 (via randomized rounding) and 4.1 (via systematical rounding) are intended to reproduce the original population. We now present some empirical evidence concerning the extent of the differences between the pseudo-populations and the original one, under both rounding approaches and with respect to a number of population parameters.

Consider 18 distinct scenarios arising by the combination of the following parameters: $N \in \{500, 1000, 5000, 10000, 20000, 30000\}$ and $f \in \{0.01, 0.05, 0.10\}$; let H, F and G (as defined in A2) be, respectively, a Normal copula with $\rho = 0.75$, a Log-Normal distribution with location parameter $\mu = 1/2$ and scale parameter $\sigma = 1/2$ ($F \sim LN(\mu = 1/2, \sigma = 1/2)$), and an Exponential distribution with parameter $\lambda = 1$ ($G \sim Exp(\lambda = 1)$). We generate a *parent* population of $N = 30000$ independent pairs from H, and consider the first 500, 1000, ..., 30000 observations to define the populations for each of the considered N values.
We investigate how well the original population is reproduced by assessing the recovery of the following population quantities: the population size, the total of the auxiliary variable, the first four moments from the origin, the median, Galton's skweness index and the Quintile Share Ratio. Details on their expressions can be found in the Appendix. Under each combination of N and f, $M = 10000$ samples are generated using a Pareto design, and the pseudo-population is constructed, with weights based on both randomized and systematical rounding. The performance of the recovery is assessed via the Monte Carlo Bias, Absolute Error and Quadratic Error as follows:

$$B = \frac{1}{M} \sum_{m=1}^{M} (\theta^* - \theta)$$

$$AE = \frac{1}{M} \sum_{m=1}^{M} |\theta^* - \theta|$$

$$QE = \frac{1}{M} \sum_{m=1}^{M} (\theta^* - \theta)^2$$

where θ denotes the true population value and θ^* is the pseudo-population one. Figures 1, 2 and 3 report the results concerning B, AE and QE, respectively.

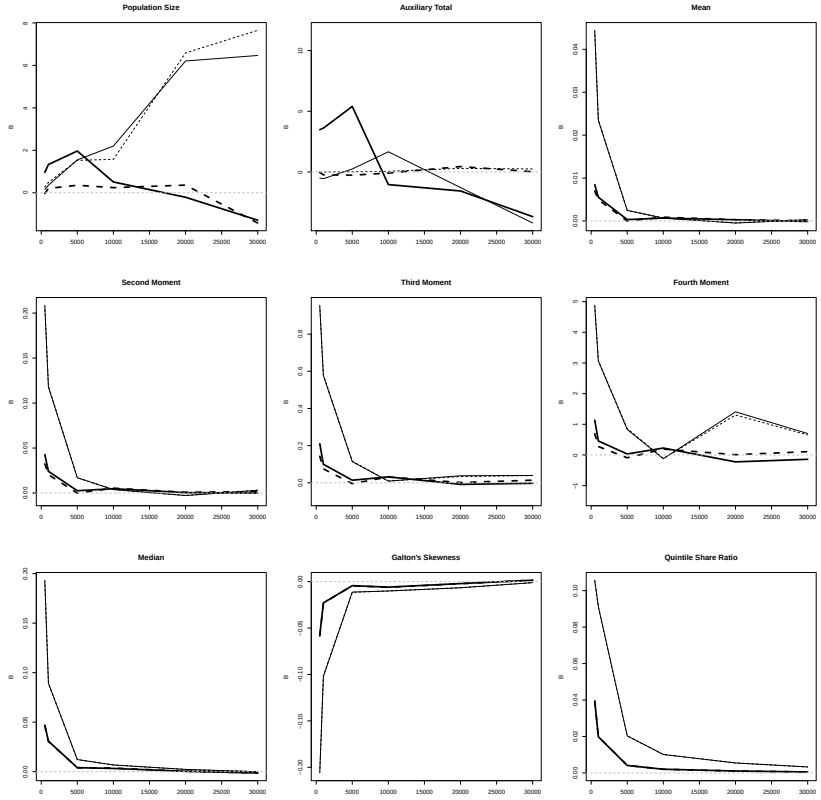

Figure 1: Monte Carlo Bias in recovering the true population parameter. The solid line refers to systematical rounding, the dashed line to randomized rounding. The thickness of the lines indicate the two levels of sampling fraction: $f = 0.01$ (thicker) and $f = 0.05$ (thinner).

Inspection of Figure 1 sheds some light on the effect of the choice of rounding method with respect to the average difference between the parameter in the pseudo-populations and their corresponding true values in the original one. The bias in recovering the population size seems to be overall limited when the sampling fraction is small ($f = 0.01$, possibly realizing the asymptotic conditions discussed in this paper); no striking differences in term of performances can be assessed between the two rounding approaches here. When considering the auxiliary total, however, we find evidence of a better mimicking via the randomized approach (dashed lines), since the bias is close to zero for all the scenarios. For all other parameters, the distortion induced by both rounding methods decreases on average as N increases and f is held fixed (faster for $f = 0.01$), and leads to biases whose magnitude is virtually indistinguishable from eachother.

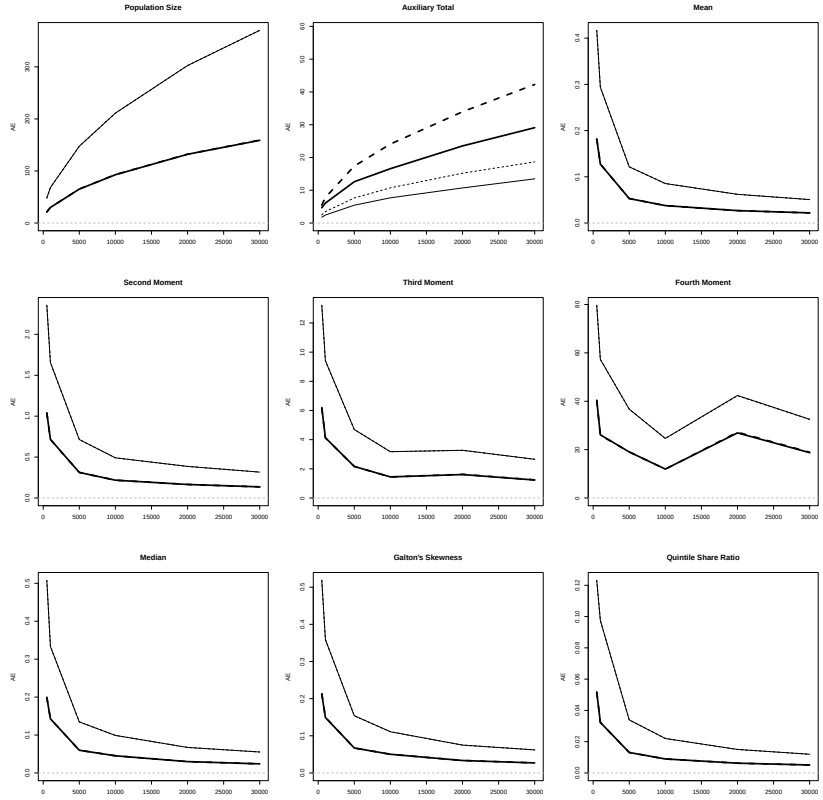

Figure 2: Monte Carlo Absolute error in recovering the true population parameter. The solid line refers to systematical rounding, the dashed line to randomized rounding. The thickness of the lines indicate the two levels of sampling fraction: $f = 0.01$ (thicker) and $f = 0.05$ (thinner).

Figure 2 reports the Monte Carlo Absolute Error in recovering the population parameters. For both population size and auxiliary total, the AE seems to be monotically increasing with N for fixed f, the worse scenarios being those with $f = 0.05$ for population size and $f = 0.01$ for auxiliary total. This could be a direct effect of the propagation error discussed in [Andreis and Mecatti (2015)], that becomes particularly evident when we consider absolute differences, where deviations of opposite sign cannot compensate. For all other parameters, the AE appears to more or less steadily decrease with N and fixed f (faster for $f = 0.01$). In all cases but for the auxiliary total, the two rounding approaches yield comparable results.

Figure 3 contains the results pertaining the Monte Carlo Quadratic Errors, that provide a measure of the variability of the recovered parameters around the true population values. Once again, the propagation of the rounding error seems to impact negatively

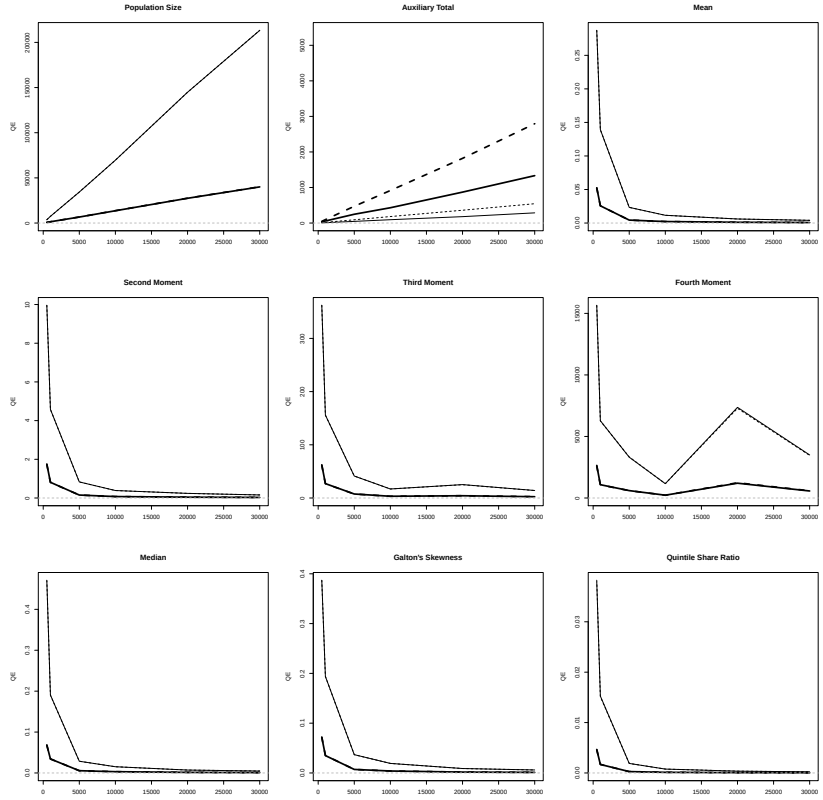

Figure 3: Monte Carlo Quadratic Error in recovering the true population parameter. The solid line refers to systematical rounding, the dashed line to randomized rounding. The thickness of the lines indicate the two levels of sampling fraction: $f = 0.01$ (thicker) and $f = 0.05$ (thinner).

on the recovery of population size and auxiliary total: their QEs increase quickly with N for fixed f, with the largest sampling fraction being worse for the population size and the smallest for the auxiliary total. In all other cases, the average quadratic error decreases with N and for fixed f (faster for $f = 0.01$). Again, apart from the auxiliary total recovery, we fail to detect significant differences imputable to the choice of the rounding method.

The comparison of the two rounding approaches with respect to recovery of the structure of the original population, motivated by the basic Bootstrap principle of mimicking, suggests an overall factual equivalence in the simulated scenarios. A possible explanation might be that the magnitude of the asymptotic difference between the two methods, stated in Proposition 4, is so limited in the cases we considered in this paper, as to be masked by Monte Carlo error. If this were indeed the case, then it might be safe

to assume that, at least under the scenarios we presented, both methods would perform equally well in practice when building pseudo-populations for resampling purposes such as the estimation of the variance of complex functionals of the survey variable. Paragraph 5.2 is concerned with providing empirical evidence on this.

5.2 Estimation of the variance of complex functionals

We now present empirical results concerning the Bootstrap estimation under both rounding approaches of the variance of some complex functionals of the survey variable that satisfy the Hadamard-differentiability condition. Specifically, we consider the estimation of the variance of the Hájek estimators of: the mean, the quartiles, Galton's skewness index and the Quintile Share Ratio. Details on the estimators can be found in the Appendix.

Consider the six distinct scenarios arising by the combination of the following parameters: $N \in \{500, 1000, 20000\}$ and $f \in \{0.01, 0.05\}$; let H, F and G be defined as in paragraph 5.1. Other combinations of shapes and distributions have been investigated, yielding comparable results. Once again, we generate a larger parent population and consider sub-sequences of $500, 1000$ and 20000 units to form the three populations from which to extract the original samples to be used to create the pseudo-populations for resampling pruposes. Given the complexity of the involved estimators, their actual variances cannot be exactly computed; we estimated them via a Monte Carlo simulation with 250000 iterations under each scenario and used such an empirical approximation as reference population value. Following [Andreis and Mecatti (2015)], we use as MC measures of performance the percentage Relative Bias ($\%RB$) and the percentage Relative Root Mean Square Error ($\%RRMSE$) of the final variance Bootstrap estimates to investigate deviations induced by the rounding practice.
We report the results of 5000 simulations entailing 500 Bootstrap samples at each run to estimate the variance of each of the estimators; we employ a Pareto sampling scheme both as original and resampling design. Tables 1 and 2 contain, respectively, the percentage relative bias and relative root mean square error under all the scenarios and for all the Bootstrap variance estimators, under both systematic and randomized rounding. All values are rounded to the second decimal place.

Inspection of Table 5.2 and 5.2 reveals that the performances of the two rounding methods are, with respect to $\%RB$ and $\%RRMSE$, essentially indistiguishable under the scenarios we consider, which is in line with what found in [Andreis and Mecatti (2015)] limited to the Horvitz-Thompson estimator, and consistent with the results discussed in paragraph 5.1, of whose scenarios these form a subset.

Table 1 clearly shows that the bias in estimating the variance of the $\widehat{\theta}_i$s decays very slowly as N increases, while holding the sampling fraction f fixed; the $\%RB$ remains in general non-negligible also with the largest population size. The estimation of the variance of the estimator of the mean is an exception and, to some extent, so is the first quartile's: the former possibly by virtue of its simple linear functional form, the latter due to the shape of the distribution of the Y, markedly left-skewed. Interestingly, at

%RB: Variance of $\widehat{\theta}_i$	$f = 0.01$			$f = 0.05$		
$\widehat{\theta}_i \setminus$ Population Size	500	1000	20000	500	1000	20000
Mean	−18.18 −18.29	−8.21 −8.32	−0.11 −0.02	−2.25 −2.57	3.66 3.71	−5.07 −4.89
First Quartile	14.28 14.34	26.98 26.83	5.17 5.19	55.28 55.03	30.28 30.35	1.36 1.26
Median	63.47 63.49	51.87 51.36	28.22 28.37	26.66 26.61	27.74 27.33	33.26 33.40
Third Quartile	49.11 48.97	128.37 127.68	58.06 57.99	131.41 130.47	104.34 104.33	54.76 55.32
Galton's Skewness	96.98 96.89	92.15 92.39	31.42 31.30	63.04 63.14	54.09 53.75	21.48 21.47
Quintile Share ratio	332.26 331.36	410.02 411.56	18.33 18.35	113.66 113.92	70.11 70.32	9.01 8.94

Table 1: Percentage Relative Bias in the estimation of the variance. For each combination, the upper value is the *%RB* under systematic rouding, while the lower refers to the randomization approach.

%RRMSE: Variance of $\widehat{\theta}_i$	$f = 0.01$			$f = 0.05$		
$\widehat{\theta}_i \setminus$ Population Size	500	1000	20000	500	1000	20000
Mean	38.11 38.02	18.62 18.66	1.03 1.04	7.64 7.54	3.95 3.97	2.04 2.03
First Quartile	49.36 49.92	30.48 30.51	2.71 2.71	18.27 18.20	9.34 9.36	2.23 2.23
Median	98.93 99.26	47.42 47.41	4.34 4.37	18.86 18.87	11.61 11.52	1.53 1.53
Third Quartile	170.44 169.98	134.05 133.18	9.60 9.60	57.79 57.34	30.83 30.85	1.90 1.90
Galton's Skewness	60.18 60.13	42.86 42.96	4.54 4.52	21.02 21.07	13.31 13.21	2.11 2.11
Quintile Share ratio	77.90 77.92	42.60 42.79	0.53 0.53	8.10 8.11	3.23 3.24	0.54 0.54

Table 2: Percentage Relative Root Mean Square Error in the estimation of the variance. For each combination, the upper value is the *%RRMSE* under systematic rouding, while the lower refers to the randomization approach.

least in the case of the mean and the second quartile, it is possible to observe an increase in *%RB* with f at the largest population size: this may be evidence of the rounding error propagation effect ([Andreis and Mecatti (2015)]). Finally, Table 2 suggests that the estimation of the variance becomes more stable as N increases, for fixed f, as indicated by a steady decrease in *%RRMSE* under all scenarios.

6 Conclusions

The present paper discussed resampling methods under general complex sampling designs, such as π-ps designs. Due to their theoretical properties, attention is focused on methodologies based on pseudo-populations. A simplified version of the pseudo-population based on systematically rounding sampling weights is often suggested in the literature for ease of implementation and for a reduced computational burden. We derived theoretical results on the extent of the risks involved in this practice, and showed that it leads to non-consistent estimation of the distribution function of the survey variable and functionals thereof. We investigated, via simulation, the practical consequences of this finding by considering two relevant issues: i) the ability of the two rounding approaches to lead to pseudo-populations that mimick well the original one, and ii) the problem of obtaining a point estimate of the variance of complex estimators, under a π-ps design. The results of the simulation study, presented in Section 5, indicate that the practice of rounding non-integer weights can have detrimental effects on the estimation of both simple and complex functionals of the distribution function regardless of the approach: the empirical evidence we have produced suggests, however, that no relevant difference can be observed under the scenarios we have considered when comparing randomized to systematical rounding in terms of both mimicking ability and final Boostrap estimates properties. In light of this and given the variety of scenarios we considered, we would recommend the use of the simplified version of rounding as a computationally lighter and non-harmful alternative to the randomized approach, at least when point estimation of the variance of a complex functional of $F(y)$ is of interest. Further numerical investigations, not shown here, concerning c and $c(y)$ (cfr., Equations (12) and (20), respectively) indicated both asymptotic quantities to be extremely small in absolute value (maximum size 10^{-5} in all considered scenarios), which is consistent with overall negligible bias in recovering population quantities, as discussed in paragraph 5.1; their magnitude may be the reason why no evidence of a difference between the two rounding method has been detected with the (finite) population sizes we considered. We reckon, however, that the construction of asymptotic interval, rather than point estimates for general functional of the distribution function based on the systematic rounding approach should lead to intervals with asymptotic coverage probability equal to zero. This will be subject to further research.

References

[Antál and Tillé (2011)] Antál, E. and Tillé, Y (2011): A direct bootstrap method for complex sampling designs from a finite population. *Journal of the American Statistical Association*, *106*, 534–543.

[Andreis and Mecatti (2015)] Andreis, F. and Mecatti, F. (2015): Rounding non-integer weights in bootstrapping non-i.i.d. samples: actual problem or harmless practice? In *Advances in Complex Data Modeling and Computational Statistics and Data Analysis*, Springer, ISBN 978-3-319-11149-0.

[Beaumont and Patak (2012)] Beaumont, J.-F. and Patak, Z. (2012): On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *International Statistical Review*, *80*(1), 127–148.

[Berger (2011)] Berger, Y.G. (2011): Asymptotic consistency under large entropy sampling designs with unequal probabilities. *Pakistan Journal of Statistics*,*27*, 407–426.

[Berger (1998)] Berger, Y.G. (1998): Rate of convergence to asymptotic variance for the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, *74*, 149–168.

[Chao and Lo (1985)] Chao, M.T. and Lo, S.H. (1985): A Bootstrap method for finite population. *Sankhyā*, A47:399–145.

[Chauvet (2007)] Chauvet, G. (2007): Méthodes de bootstrap en population finie. PhD Dissertation, Laboratoire de statistique d'enquêtes, CREST-ENSAI, Université de Rennes 2. Available at http://tel.archives-ouvertes.fr/docs/00/26/76/89/PDF/thesechauvet.pdf.

[Conti (2014)] Conti, P.L. (2014): On the estimation of the distribution function of a finite population under high entropy sampling designs, with applications. To appear in *Sankhyā* B, DOI:10.1007/s13571-014-0083-x.

[Conti et al. (2015)] Conti, P.L., Marella, D., Mecatti, F. (2015): Recovering sampling distributions of statistics of finite populations via resampling: a predictive approach. *Submitted for publication* (available upon request from the authors).

[Csörgő and Rosalsky (2003)] Csörgő, S. and Rosalsky, A. (2003): A survey of limit laws for bootstrapped sums. *International Journal of Mathematics and Mathematical Sciences*, *45*, 2835–2861.

[Gross (1980)] Gross, S.T. (1980): Median estimation in sample surveys. In *ASA Proceedings of the Section on Survey Research Methods*, 181–181. American Statistical Association.

[Hájek (1964)] Hájek, J. (1964): Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, *35*, 1491–1523.

[Hájek (1981)] Hájek, J. (1981): Sampling from a Finite Population (Statistics: Textbooks & Monographs). Dekker New York.

[Hall (1992)] Hall, P. (1992): The Bootstrap and Edgeworth Expansion. Springer Series in Statistics.

[Holmberg (1998)] Holmberg, A. (1998): A bootstrap approach to probability proportional to size sampling. In *Proceedings of Section on Survery Research Methods*. American Statistical Association, 181–184.

[Ranalli and Mecatti (2012)] Ranalli, M.G. and Mecatti, F. (2012): Comparing recent approaches for bootstrapping sample survey data: a first step towards a unified approach. In *Proceedings of Section on Survey Research Methods*. American Statistical Association, 4088–2099.

[R Core Team (2015)] R Core Team (2015): R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, url: http://www.R-project.org*.

[Serfling (1980)] Serfling, R.J. (1980): Approximation Theorems of Mathematical Statistics. Wiley Series in Probability and Statistics. John Wiley & Sons.

[Tillé (2006)] Tillé, Y. (2006): Sampling Algorithms. Springer Series in Statistics. Springer New York.

[van der Vaart (1998)] van der Vaart, A.W. (1998): Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

7 Appendix

7.1 Population quantities used in paragraph 5.1

Expressions for the population quantities used to compare the pseudo-populations to the original ones.

$$
\begin{aligned}
\text{Population Size: } N &= \sum_{i=1}^{N} 1 \\
\text{Auxiliary Total: } X &= \sum_{i=1}^{N} x_k \\
\text{Moments: } \mu_j &= \frac{1}{N}\sum_{i=1}^{N} y_i^j, j = 1,2,3,4 \\
\text{Quantiles: } Q(p) &= \inf\{y : F(y) \geq p\} \\
\text{Galton's Skewness: } \lambda &= \frac{Q(0.75) + Q(0.25) - 2Q(0.50)}{Q(0.75) - Q(0.25)} \\
\text{Quintile Share Ratio: } QSR &= \frac{\sum_{i=1}^{N} y_i I_{(y_i \leq Q(0.20))}}{\sum_{i=1}^{N} y_i I_{(y_i \geq Q(0.80))}}
\end{aligned}
$$

7.2 Estimators used in paragraph 5.2

Expressions for the Hájek estimators considered in Section 5:

$$\begin{aligned}
\text{Mean: } \widehat{\mu}_H &= \frac{\sum_{i\in\mathbf{s}} \pi_i^{-1} y_i}{\sum_{i\in\mathbf{s}} \pi_i^{-1}} \\
\text{Quantiles: } \widehat{Q}_H(p) &= \inf\{y : \widehat{F}_H(y) \geq p\} \\
\text{Galton's Skewness: } \widehat{\lambda}_H &= \frac{\widehat{Q}_H(0.75) + \widehat{Q}_H(0.25) - 2\widehat{Q}_H(0.50)}{\widehat{Q}_H(0.75) - \widehat{Q}_H(0.25)} \\
\text{Quintile Share Ratio: } \widehat{QSR}_H &= \frac{\sum_{i\in\mathbf{s}} \pi_i^{-1} y_i I_{(y_i \leq \widehat{Q}_H(0.20))}}{\sum_{i\in\mathbf{s}} \pi_i^{-1} y_i I_{(y_i \geq \widehat{Q}_H(0.80))}}
\end{aligned}$$

The definition of the estimator of the distribution function $\widehat{F}_H(y)$ is in Section 2.1.